DOCUMENT RESUME

ED 299 281 TM 011 942

AUTHOR Tollefson, Nona; Chung, Jing-Mei
TITLE A Comparison of Two Methods of Assessing Partial Knowledge on Multiple-Choice Tests.
SPONS AGENCY Kansas Univ., Lawrence. General Research Fund.
PUB DATE [86]
NOTE 17p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; *Comparative Analysis; Evaluation Methods; Graduate Students; Guessing (Tests); Higher Education; *Multiple Choice Tests; *Reliability; *Scoring Formulas
IDENTIFIERS Correction for Guessing; Elimination Scoring; Internal Consistency; Number Right Scoring; *Partial Knowledge (Tests)

ABSTRACT

Procedures for correcting for guessing and for assessing partial knowledge (correction-for-guessing, three-decision scoring, elimination/inclusion scoring, and confidence or probabilistic scoring) are discussed. Mean scores and internal consistency reliability estimates were compared across three administration and scoring procedures for multiple-choice tests: (1) number right; (2) Coombs elimination response method; and (3) inclusion. It was hypothesized that the mean scores of subjects administered an achievement test under experimental inclusion-of-correct- and elimination-of-incorrect-alternatives directives would be significantly different from the mean score of subjects administered the same test under number-right directives. Seventy-five graduate students in an introductory statistics class were randomly assigned to one of three administration and scoring procedures. A 36-item multiple-choice test with 4-response alternatives per item was used to assess the effect of the different methods. A comparison of mean scores indicated no significant between-group effect. Elimination and inclusion scoring produced higher internal consistency reliability estimates than did number-right scoring. Two data tables conclude the document. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED299281

A Comparison of Two Methods of Assessing Partial Knowledge on Multiple-Choice Tests

Nona Tollefson and Jing-Mei Chung
The University of Kansas

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

NONA TOLLEFSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM011942

This study compared mean scores and internal consistency reliability estimates across three administration and scoring procedures: number right, elimination, and inclusion. Students in an introductory statistics class were randomly assigned to one of three administration and scoring procedures. A comparison of mean scores indicated no significant between-group effect. Elimination and inclusion scoring produced higher internal consistency reliability estimates than did number-right scoring.

The general acceptance of the multiple-choice test item as the best format for objective measurement of aptitude or achievement does not mean the item format does not have limitations; it does not control for guessing, and it does not assess level of partial information. On a conventional multiple-choice test item, students evaluate each option and choose the one that is most correct. Students may hold a number of beliefs about these options: one of the options may appear to be clearly correct and the others clearly incorrect, one or two options may appear correct and the remainder clearly incorrect, or all of the options may appear to have some chance of being correct. However, students are not allowed to describe their analysis of these options under conventional test directions but must select a single option. By allowing students to select only one option, an instructor is losing information about their degree of knowledge, specifically their level of partial information.

The conventional number-right method (NR) of scoring multiple-choice items awards one point when the correct answer is chosen and no points when a distractor is selected as the correct response. Complete knowledge or a lucky guess leads to an item score of one and complete misinformation to a score of zero. Partial knowledge may lead to a score of either one or zero. The inability of NR scoring to discriminate between partial information, complete information/misinformation, or a lucky guess is a disadvantage of this scoring method.

This paper was funded inpart by a grant from the General Research Fund at the University of Kansas.

A number of studies have called attention to the advantages of considering the examinee's evaluation of wrong options in multiple-choice item scoring. When an examinee fails to select the correct option, his/her rejection of the correct option and his/her choice of an incorrect option have a significance that ought to be recognized in scoring the examination. Both logical and experimental findings indicate the usefulness of introducing this degree of sophistication into multiple-choice test scoring; failure to do so has been identified as a major weakness of the multiple-choice item (Highet, 1950).

Procedures for correcting for guessing and for assessing partial knowledge have been proposed over the years. These procedures include: correction-for-guessing (CFG), three-decision scoring, elimination/inclusion scoring, and confidence or probabilistic scoring. All are based on a concept of an item "response" that is broader than simply indicating the correct alternative (Wang & Stanley, 1970). Among measurement specialists, considerable optimism, regarding improved reliability and validity, remains for these procedures (see, for example, Collet, 1971; Hambleton, Roberts, & Traub, 1970, Jaradat & Sawaged, 1986).

Correction-for-Guessing Scoring Method

Since the beginning of the objective-test movement, there has been concern over the effects of guessing on the reliability and validity of multiple-choice tests, and testers have attempted to correct for guessing. However, CFG scoring methods are based on the unrealistic assumption that, if a person does not know the answer to a question, s/he guesses randomly among the response options. The differential popularity of incorrect options belies this assumption. Moreover, Sax and Collet (1969) have demonstrated that instructions not to guess can be as effective as formula scoring in correcting for guessing, and Salkter (1968) has shown that instructions that accompany formula-scored tests tend to penalize lov-risk takers.

A second weakness of CFG scoring is that it fails to account for partial information/misinformation. Differential weighting of distractors (assigning weights based on degree of correctness) represents one approach to identifying partial information/misinformation. Empirical methods are available for deriving weights; however, they do not ensure that the more heavily weighted options are necessarily more correct. Thus, although the correctness of an option is defined operationally, it is the response options rather than the examinees' pattern of selecting correct/incorrect responses that bear the burden of differentiating among examinees with different levels of partial knowledge.

Coombs Elimination Response Method

Coombs, Milholland and Womer (1956) developed a scoring method in which respondents were instructed to eliminate incorrect alternatives (ES). If r of k alternatives were eliminated, the score was +r if the correct response was not eliminated and r - k if it was. Thus a four alternative multiple-choice item yielded scores from -3 to +3, a seven point scale, instead of the usual 0 or 1 scores (Coombs, 1953).

Dressel and Schmid (1953), using achievement tests, examined an early variant of ES scoring, referred to as the "free-choice" method. Examinees marked as many options as needed to ensure that the correct answer was not omitted. Each incorrect alternative marked was assigned a score of -1/4, with the correct option scored +1. This scoring method resulted in an insignificant gain in internal consistency, compared to NR scoring; validity was not examined.

Collet (1971), using IQ tests, compared three scoring approaches: ES, NR scores corrected for guessing, and an empirically determined weighted-choice method. He found no significant gain in reliability or criterion-related validity for ES compared to NR scoring.

Hakstian and Kansup (1975) compared ES and NR scoring. Their results indicated that ES did not significantly increased reliability or validity, and they concluded that there was no reason to prefer ES to NR scoring. Traub and Fisher (1977) also compared ES and NR scoring. No differences in factorial structure were found between tests administered and scored under ES and NR formats. Although reliabilities were not the central focus of Traub and Fisher's study, reliabilities for ES were generally equal to or greater than reliabilities for NR. Jaradat and Sawaged (1986) compared inclusion scoring (IS) (Dressel & Schmid, 1953), ES, and NR scoring and found IS scores were more reliable than NR scores and more valid than NR and CFG scores.

Three-Decision Multiple-Choice Tests

The three-decision multiple-choice test (Willey, 1960) has been proposed as an improvement over the conventional multiple-choice test, because it enables the examinee to earn partial credit for partial knowledge. This method requires the examinee to designate two alternatives which s/he feels are definitely wrong as well as designating the alternative which s/he feels is the correct one. Three points are awarded if the alternative designated as the correct one is actually the correct alternative. Two points are awarded, if the alternative which is actually correct is not designated as the correct alternative or one of the definitely wrong alternatives. No points are awarded, if an alternative designated as one of the definitely wrong alternatives is actually the correct alternative.

Willey (1960) reported that examinees regarded the three-decision test format as fairer than NR scoring, because the influence of chance was reduced and they received partial credit for partial knowledge. The major advantage of the three-decision test is that correction for guessing is built into the test. For example, a student who did not identify 10 correct responses but who avoided

calling the correct answer "definitely wrong," would have a penalty score of 10. Another student who did not identify the same 10 correct answers, but selected incorrect responses at random would have a penalty score of 20.

Bernhardson (1967) compared three-decision and conventional NR scoring methods. Two multiple-choice tests, one with five alternatives and one with four alternatives were scored using a three-decision format and a NR format. The different scoring systems were evaluated by correlating test scores with the average mark each student obtained in all courses during the year; no significant difference in coefficients was found.

Confidence, or Probabilistic Testing Method

Confidence or probabilistic scoring methods require the examinee to indicate the degree of confidence, or subjective probability that an option (or in some cases, each option) is correct. Dressel and Schmid (1953) employed a "degree of certainty" response procedure with achievement tests, in which examinees expressed their confidence, on a scale of 1 to 4, that each response was a correct response. An insignificant gain in internal consistency reliability over NR scoring was observed. Hopkins, Hakstian, and Hopkins (1973) employed a similar response method using a 3-point scale and found a slight, but nonsignificant, increase in internal consistency reliability for confidence-weighted responses, and a slight, but nonsignificant, decrease in validity. Using a 10-point confidence distribution method, Michael (1968) obtained a gain in internal consistency reliability for the confidence scores, but the gain was not assessed for statistical significance.

Using more sophisticated confidence-responding procedures and scoring functions, Rippey (1968), Hambleton et al. (1970), and Koehler (1971) found no gain in internal consistency reliability for confidence scoring compared to NR

scoring. In the latter two studies (both employed achievement tests), validity coefficients were also examined, and no significant increases in validity were obtained when the confidence-weighting procedures were used.

The Purpose of the Study

The primary purpose of this study was to compare mean scores and internal consistency reliability estimates produced by NR, IS, and ES scoring. It was hypothesized that the mean scores of the groups administered an achievement test under experimental IS and ES directions would be significantly different from the mean score of the group administered the same test under NR directions. For the same reason, it was expected that test scores obtained under either of the two experimental methods would be more reliable than scores obtained from the NR method.

Subjects

The sample included 75 graduate students enrolled in two sections of a statistics course. These students were randomly assigned to one of three groups: ES, IS, or NR. All students completed the same test; only the administration and scoring procedures varied by group.

Instrument

A 36 item multiple-choice test with 4-response alternatives per item was used to assess the effect of different administration and scoring methods. The major content areas covered on this test were: basic probability and descriptive statistics; discrete and continuous random variables; and sampling distributions and estimation. All items had been field-tested, and item difficulties were in the moderate (.40 to .60) to easy (.80 to 1.0) difficulty range.

Administration and Scoring Procedures

Students were administered the test under testing conditions. A general instruction period was provided, and students were informed that three different forms of the test were being administered. When a question arose, the examiner went to the desk of the student and quietly answered the question, so as not to confuse students completing the test under different sets of directions.

NR scoring gave one point when the correct answer was chosen and zero when a distractor was chosen. Total scores for the NR condition could range from 0 to 36 on the 36-item test.

IS scoring awarded 3 points for selecting the correct answer and extracted a penalty of 1 point for identifying each incorrect answer as a correct one. Scores on a single question ranged between -3 and +3. Consequently, test scores could vary from -108 to +108 on the 36-item test.

ES scoring awarded 1 point for identifying each incorrect answer and a penalty of three points for identifying the correct answer as an incorrect answer. Scores on a single question ranged between -3 and +3, and total test scores could range from -108 to +108.

Simulated NR Scoring Procedure. In order to simulate scores obtained on a NR test, the following procedure was employed:

For the IS group, if a person marked only one option and it was correct, s/he received a score of 1; otherwise, a score of zero was given. The author called it NR(IS).

For the ES group, if a person eliminated all three incorrect options, s/he received a score of 1; otherwise, a score of zero was given. The author called it NR(ES).

The simulated item scores were summed over all items. On the 36-item test this scoring procedure yielded a range of scores from 0 to 36. Coombs et al.

(1956) calculated a similar score and called it the theoretical conventional score. Bradbard and Green (1985) used a similar procedure called simulated standard test procedure.

Data Analysis

Data were analyzed using mean difference and correlational procedures. Four hypotheses were tested.

(1) There is no significant difference in mean scores under IS and ES directions.

(2) There are no significant differences in mean scores under NR, NR(IS), and NR(ES) scoring.

(3) There is no significant difference in the relationship between IS and NR(IS) and ES and NR(ES).

(4) There is no significant differences in the internal consistency reliability coefficients under the three scoring methods.

Results

Means and standard deviations for the three scoring methods and for the stimulated NR scoring methods are presented in Table 1. A *t* test was used to compare mean scores under IS and ES. This analysis yielded a nonsignificant difference in means (t (48) = .08). A one-way analysis of variance was used to compare the mean scores for NR, NR(IS), and NR(ES). This analysis yielded a nonsignificant *F* (F (2, 74) = 1.70).

Insert Table 1 Here

Scores for IS and NR(IS) and for ES and NR(ES) were correlated using the Pearson Product-Moment Correlation Coefficient. IS and NR(IS) were correlated .84 and ES and NR(ES) were correlated .94. These coefficients were compared using the two sample z test. This analysis yielded a nonsignificant z value ($z = -1.81$).

Finally, internal consistency reliabilities were computed using coefficient alpha for the three scoring methods. These coefficients are reported in Table 2. Each pair of coefficients was compared using the z test for independent samples. The results of this analysis, also reported in Table 2, yielded nonsignificant differences in reliability coefficients.

Insert Table 2 Here

Discussion

The results of this study failed to demonstrate that scoring methods designed to assess level of partial information produced scores that differed from NR scores in either average performance or reliability. Several factors may explain the absence of significant results. First, and probably most important, the test items were relatively easy, and most students had no difficulty selecting the correct answer. Stated differently, students' knowledge level was high, and they may well have had complete information for questions of the difficulty included on the test. A longer, more difficult test would produce higher reliabilities and perhaps differential increases in test reliability for ES and IS. Finally, students in the experimental conditions did not utilize the opportunity to indicate different levels of information.

Students administered the test under IS scoring typically selected one response option, while students administered the test under ES scoring typically selected three incorrect options. Students' failure to exercise the option of selecting more than one possibly correct option may have reflected either a high level of knowledge or a reluctance to employ a new test-taking strategy under actual testing conditions.

In spite of the nonsignificant findings, it seems reasonable to continue to explore administration and scoring methods that permit assessment of students' levels of information. This study, like other before it, found an increase in reliability associated with IS and ES scoring. Future studies will need to focus on the impact of partial credit scoring methods on test reliability for tests of different levels of difficulty.

Finally, teachers' and students' responses to these test administration and scoring method need to be explored. Bradbard and Green (1985) report that instructors like the Coombs' procedure (ES), because it "seems to provide finer discriminations among students." On the other hand, tests must be scored by hand, and machine-scoring can not be used. Jaradat and Tollefson (in press) report that students find the ES and IS testing procedures fair, but complain about their difficulty. In summary, instructors and students report some problems adjusting to the ES and IS methods of responding principally, because they are perceived as more difficult than the conventional method of responding. However, this disadvantage may be offset in the instructor's mind by the perceived greater discrimination provided by ES or IS and in the student's mind by the belief that they are fairer than NR scoring.

Table 1

Cell Means and Standard Deviation for the Three Scoring Methods

Source of Variation	Means	Standard Deviation
IS	79.96	14.16
ES	80.28	14.68
NR	29.68	3.22
NR(IS)*	28.28	3.22
NR(ES)**	28.08	3.57

*NR(IS) ------ simulated NR scores from IS scores

**NR(ES) ------ simulated NR scores from ES scores

Table 2

Coefficient Alpha Reliability Coefficients for the Three Scoring Methods

Groups	Alpha		z
NR	0.4673	NR vs NR(IS)	.13
NR(IS)	0.5009	NR vs NR(ES)	.61
NR(ES)	0.6004	NR(IS) vs NR(ES)	.48

REFERENCES

Abu-Sayf, F. K. (1975). Relative effectiveness of the conventional formula score, Journal of Educational Research, 69, 160-162.

Abu-Sayf, F. K. (1977). A new formula score, Educational and Psychological Measurement, 7, 853-862.

Abu-Sayf, F. K. (1979). Recent developments in the scoring of multiple-choice items, Educational Review, 31, 269-279.

Abu-Sayf, F. K., & Diamond, J. J. (1976). Effect of confidence level in multiple-choice test answers on reliability and validity of scores, Journal of Educational Research, 70, 62-63.

Aiken, L. R. (1968). Weighting and guessing on varieties of the multiple-choice item, Educational and Psychological Measurement, 28, 1087-1101.

Aiken, L. R. (1979). Scoring for partial knowledge on the generalized rearrangement item, Educational and Psychological Measurement, 30, 87-94.

Bernhardson, C. S. (1967). Comparison of the three-decision and conventional multiple-choice tests, Psychological Reports, 20, 695-698.

Bradbard, D. A., & Green, S. B. (1985). Use of the Coombs elimination procedure in classroom tests, Journal of Experimental Education, 54, 2, 68-72.

Collet, L. S. (1971). Elimination scoring: An empirical evaluation, Journal of Educational Measurement, 8, 209-214.

Coombs, C. H. (1953). On the use of objective examination, Educational and Psychological Measurement, 13, 308-310.

Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge, Educational and Psychological Measurement, 16, 13-37.

Dressel, P. L., & Schmid, P. (1953). Some modifications of the multiple-choice item, Educational and Psychological Measurement, 13, 574-595.

Hakstian, A. R., & Kansup, W. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: II. Testing procedures, Journal of Educational Measurement, 12, 231-239.

Hambleton, R. K., Roberts, D. M., & Traub, R. E. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test, Journal of Educational Measurement, 7, 75-82.

Highet, G. (1950). The Art of Teaching. New York: Knopf.

Hopkins, K. D., Hakstian, A. R., & Hopkins, B. R. (1973). Validity and reliability consequences of confidence weighting, Educational and Psychological Measurement, 33, 135-141.

Jaradat, D., & Sawaged, S. (1986). The Subset selection technique for multiple-choice tests: An empirical inquiry. Education and Psychological Measurement, 23, 369-376.

Jaradat, D. & Tollefson, N. (In press). The impact of alternative scoring procedures for multiple-choice items on test reliability, validity, and grading. Educational and Psychological Measurement.

Kansup, W., & Hakstian, A. R. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures, Journal of Educational Measurement, 12, 219-230.

Koehler, R. A. (1971). A comparison of the validities of conventional choice testing and various confidence marking procedures, Journal of Educational Measurement, 8, 297-303.

Lord, F. M. (1975). Formula scoring and number-right scoring, Journal of Educational Measurement, 12, 7-11.

Michael, J. C. (1968). The reliability of a multiple-choice examination under various test-taking instructions, Journal of Educational Measurement, 5, 307-314.

Rippey, R. M. (1968). Probabilistic testing, Journal of Educational Measurement, 5, 211-215.

Rippey, R. M. (1970). A comparison of five different scoring functions for confidence tests, Journal of Educational Measurement, 7, 165-170.

Sax, G., & Collet, L. (1969). The effect of differing instructions and guessing formulas on reliability and validity, Educational and Psychological Measurement, 29, 665-680.

Slakter, M. J. (1968). The penalty for not guessing, Journal of Educational Measurement, 5, 141-144.

Traub, R. E., & Fisher, C.W. (1977). On the equivalence of constructed-response and multiple-choice tests, Applied Psychological Measurement, 1, 355-369.

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies, Review of Educational Research, 40, 663-705.

Willey, C. F. (1960). The three-decision multiple-choice test: a method of increasing the sensitivity of the multiple-choice item, Psychological Reports, 7, 475-477.